



ИНФОРМАТИКА

СПО

КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРЕ

КЛЮЧЕВЫЕ СЛОВА

- ◆ текстовая информация
- ◆ кодирование
- ◆ кодовые таблицы

КОМПЬЮТЕРНОЕ ПРЕДСТАВЛЕНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

Для компьютерного представления текстовой информации достаточно:



...	...
64	01000000
65	01000001
66	01000010
67	01000011
68	01000100

Определить алфавит
(множество всех
символов)

Присвоить каждому
символу алфавита
порядковый номер

Перевести номер
символа в двоичную
систему счисления



КОДИРОВКА ASCII

American Standard Code for Information Interchange – американский стандартный код для обмена информацией, разработанный в 1960-х годах в США.

0	0	0	0	0	0	0	0	0	1	5						
0	NUL	SOH	STX	ETX	EOT	ENC										
1	0	0	1	0	0	0	0	0								
2																
3	0															
4	@	A														
5	P															
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

Изображаемые символы
(буквы латинского алфавита, цифры, знаки препинания и арифметических операций, скобки и некоторые специальные символы)

Первые 32 символа и 128-й – управляющие
(при выводе текста они не отображаются графически)

0 1 1 1 1 1 1 0

РАСШИРЕНИЕ КОДИРОВКИ ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
	0 0 0 0 0 0 0 0 5															
0	NUL	SOH	STX	ETX	EOT	ENC										
1	DLE	DC1	DC2	DC3	DC4	NAK										
2		!	“	#	\$	%										
3	0	1	2	3	4	5										
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	1 0 0 0 0 0 0 0							u	v	w	x	y	z			
8	Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ
9	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ
A	Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ
B	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ
C	Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ
D	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ
E	Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ
F	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ	Ѷ	ѷ	Ѹ	ѹ	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ

Стандартная часть кода (0 ... 127)

КОИ-8

Расширение ASCII (128 ... 255)
 (буквы национального алфавита,
 символы национальной валюты и т.п.)

РАСШИРЕНИЕ КОДИРОВКИ ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	“	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	Windows-1251					КОИ-8				
8	Ђ	Ѓ	И	Ј	Љ	Њ	Ќ	Ў	Ѕ	Ї	Љ	Њ	Ќ	Ў	Ѕ	Ї
9	ђ	ѓ	и	ј	љ	њ	ќ	у	ѕ	ѝ	љ	њ	ќ	у	ѕ	ѝ
A	Ѐ	Ђ	Ѓ	Д	Е	Ж	З	И	Ј	Љ	Њ	Ќ	Ў	Ѕ	Ї	Љ
B	°	±	І	İ	Г	М	П	·	ё	№	€	»	Ј	Ѕ	Ѓ	İ
C	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
D	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
E	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
F	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я



СТАНДАРТ UNICODE

Unicode — это «уникальный код для любого символа, независимо от платформы, независимо от программы, независимо от языка» (www.unicode.org).

Стандарт Unicode был разработан в 1991 году и описывает алфавиты всех известных, в том числе и «мертвых», языков. Для языков, имеющих несколько алфавитов или вариантов написания (японского и индийского), закодированы все варианты.

В кодировку Unicode внесены все математические и иные научные символьные обозначения и даже некоторые придуманные языки (язык эльфов из трилогии Дж. Р. Р. Толкина «Властелин колец»).



КЛАВИАТУРЫ НЕКОТОРЫХ СТРАН МИРА



РУССКАЯ



АМЕРИКАНСКАЯ



АРАБСКАЯ



АРМЯНСКАЯ



ЯПОНСКАЯ

КОДИРОВКИ СТАНДАРТА UNICODE

Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

Для представления символов в памяти компьютера в стандарте Unicode имеется несколько кодировок.

Кодировка UTF-16



Часто используемые символы:
2 байта (16 бит)

Редко используемые символы:
4 байта (32 бит)

Кодировка UTF-8



Символы, входящие в таблицу ASCII: *1 байт (8 бит)*

Символы, не входящие в таблицу ASCII: *2-4 байта (16-32 бит)*

ИНФОРМАЦИОННЫЙ ОБЪЕМ СООБЩЕНИЯ

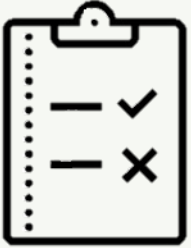
Информационным объёмом текстового сообщения называется количество бит (байт, килобайт, мегабайт и т. д.), необходимых для записи этого сообщения путём заранее оговоренного способа двоичного кодирования.

Количество символов
в сообщении

$$I = K \cdot i$$

ASCII, KOI-8, Windows-1251, ...
1 символ = 1 байт

Unicode
1 символ = 2 байта



ПРИМЕР 1

Оценим в байтах объём текстовой информации в современном словаре иностранных слов из 740 страниц, если на одной странице размещается в среднем 60 строк по 80 символов (включая пробелы).

Дано:

$i = 1$ байт

$K = 80 \cdot 60 \cdot 740$

$I = ?$

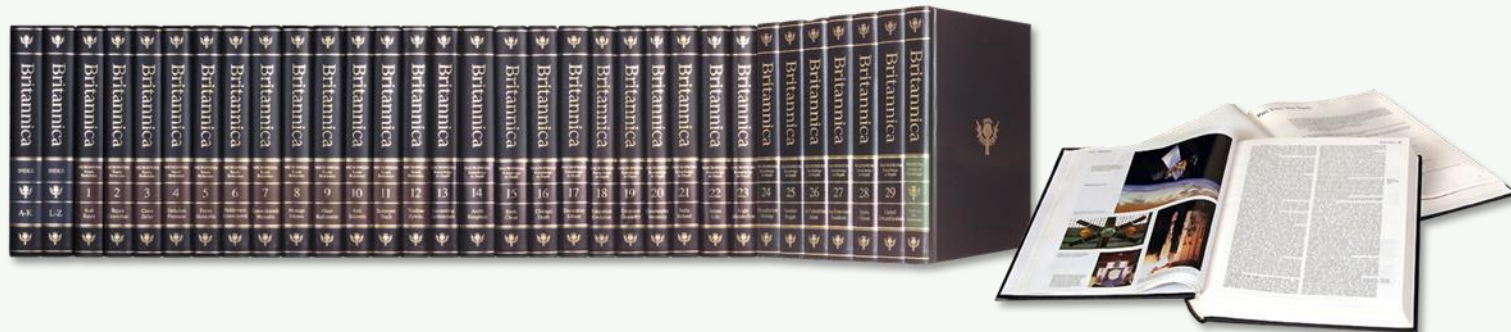
$I = K \cdot i$

$$I = \frac{80 \cdot 60 \cdot 740}{1024 \cdot 1024} \text{ Мб} \approx 3,39 \text{ Мбайт}$$

Ответ: 3,39 Мбайт

Если же использовать кодировку UTF-16, то объём этой же текстовой информации в байтах возрастет в 2 раза и составит 6,78 Мбайт.

ПРИМЕР 2



В 15-м издании энциклопедии Britannica 32 тома, в каждом из которых порядка 1000 страниц. На одной странице размещается в среднем 70 строк по 120 символов (включая пробелы) в каждой. Найдите объем текстовой информации в энциклопедии, если при записи используется кодировка Unicode («*один символ — два байта*»).

Дано:

$i = 2$ байта

$K = 32 \cdot 1000 \cdot 70 \cdot 120$

$I = ?$

$I = K \cdot i$

$$I = \frac{32 \cdot 1000 \cdot 70 \cdot 120 \cdot 2}{1024 \cdot 1024} \text{ Мб} \approx 513 \text{ Мб}$$

Ответ: 513 Мбайт

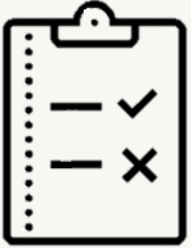


Текстовая информация по своей природе дискретна, так как представляется последовательностью отдельных символов.

В памяти компьютера хранятся специальные кодовые таблицы, в которых для каждого символа указан его двоичный код. Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

Основой для компьютерных стандартов кодирования символов послужил код ASCII, рассчитанный на передачу только английского текста. Расширения ASCII — кодировки, в которых первые 128 символов кодовой таблицы совпадают с кодировкой ASCII, а остальные (со 128-го по 255-й) используются для кодирования букв национального алфавита, символов национальной валюты и т. п.

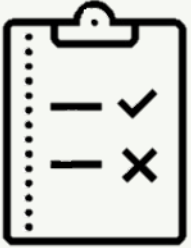
В 1991 году был разработан новый стандарт кодирования символов, получивший название Unicode (Юникод), позволяющий использовать в текстах любые символы любых языков мира. Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.



ВОПРОСЫ И ЗАДАНИЯ

Какова основная идея представления текстовой информации в компьютере?

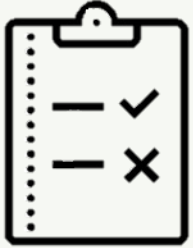




ВОПРОСЫ И ЗАДАНИЯ

Что представляет собой кодировка ASCII? Сколько символов она включает? Какие это символы?

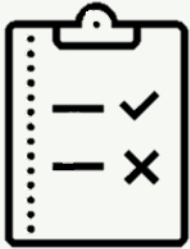




ВОПРОСЫ И ЗАДАНИЯ

Как известно, кодовые таблицы каждому символу алфавита ставят в соответствие его двоичный код. Как в таком случае вы можете объяснить вид таблицы 3.8 «Кодировка ASCII»?





ВОПРОСЫ И ЗАДАНИЯ

С помощью таблицы

а) декодируйте сообщение

64 65 73 6B 74 6F 70;

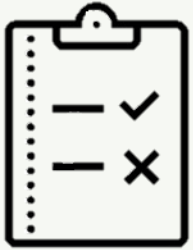
б) запишите в двоичном коде

сообщение TOWER;

в) декодируйте сообщение

01101100 01100001 01110000 01110100 01101111 01110000

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL



ВОПРОСЫ И ЗАДАНИЯ

Что представляют собой расширения ASCII-кодировки?
Назовите основные расширения ASCII-кодировки, содержащие русские буквы.

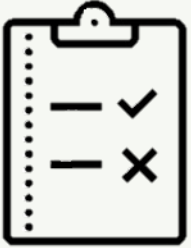




ВОПРОСЫ И ЗАДАНИЯ

Сравните подходы к расположению русских букв в кодировках Windows-1251 и КОИ-8.





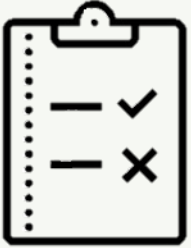
ВОПРОСЫ И ЗАДАНИЯ

Представьте в кодировке Windows-1251 текст

Знание — сила!

- 1) шестнадцатеричным кодом;
- 2) двоичным кодом;
- 3) десятичным кодом.





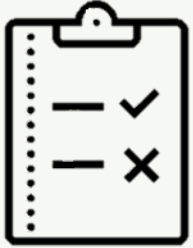
ВОПРОСЫ И ЗАДАНИЯ

Представьте в кодировке КОИ-8 текст

Дело в шляпе!

- 1) шестнадцатеричным кодом;
- 2) двоичным кодом;
- 3) десятичным кодом.





ВОПРОСЫ И ЗАДАНИЯ

Что является содержимым файла, созданного в современном текстовом процессоре?





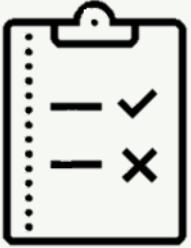
ВОПРОСЫ И ЗАДАНИЯ

В кодировке Unicode на каждый символ отводится 2 байт.

Определите в этой кодировке информационный объём следующей строки:

Где родился, там и содился.

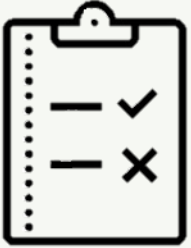




ВОПРОСЫ И ЗАДАНИЯ

Набранный на компьютере текст содержит 2 страницы. На каждой странице 32 строки, в каждой строке 64 символа. Определите информационный объём текста в кодировке Unicode, в которой каждый символ кодируется 16 битами.

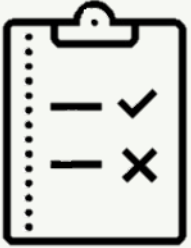




ВОПРОСЫ И ЗАДАНИЯ

Текст на русском языке, первоначально записанный в 8-битном коде Windows, был перекодирован в 16-битную кодировку Unicode. Известно, что этот текст был распечатан на 128 страницах, каждая из которых содержала 32 строки по 64 символа в каждой строке. Каков информационный объём этого текста?





ВОПРОСЫ И ЗАДАНИЯ

Определите информационный объём сообщения

`Pascal, Python, JavaScript, Java, Ruby, C++,
Kotlin, Swift, Go` – языки программирования.

- 1) в 8-битной кодировке КОИ-8;
- 2) в 16-битной кодировке UNICODE;
- 3) в кодировке UTF-8.

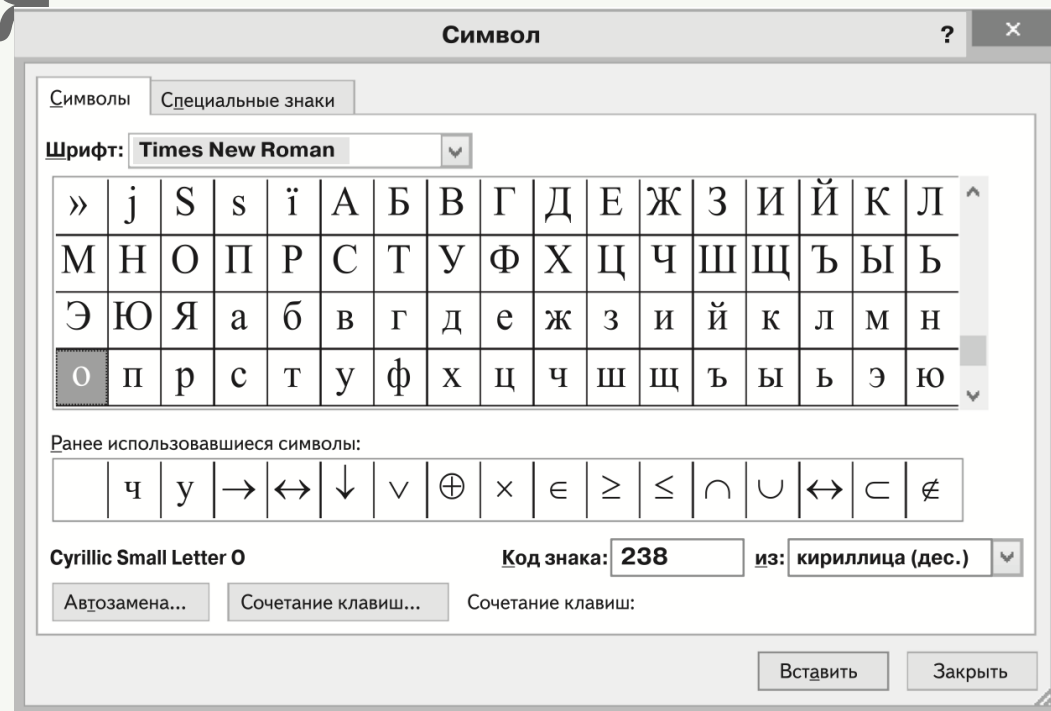




ВОПРОСЫ И ЗАДАНИЯ

В текстовом процессоре Microsoft Word откройте таблицу символов (**Вставка** → **Символ** → **Другие символы**).

В поле **Шрифт:** установите **Times New Roman**, в поле **из** — **кириллица (дес.)**. Вводя в поле **Код знака** десятичные коды символов, декодируйте сообщение:



196	238	240	238	227	243	32
238	241	232	235	232	242	32
232	228	243	249	232	233	46

